ISSN 2816-8089

RESEARCH ARTICLE

Prognostic Analysis of Machine Learning Techniques for Breast Cancer

Pavan Kumar SP[1], Samiha CM[2], Gururaj HL[1], Ram Kumar K[3]

[1]*Department of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, India*

[2] *Mysore Institute of Commerce and Arts, Mysuru, India*

[3]*Department of Statistics, Vishwakarma University, Pune, India*

Abstract

Breast cancer is the second most common cause of cancer death in women, after lung cancer. Mutations in genes regulate cell development, and mutations proliferate and divide cells in an uncontrolled manner, resulting in cancer. There are five stages of breast cancer. In each stage, the size of the tumor varies. Alcohol consumption, body weight, history of breast cancer, age, genetics, hormone treatments, etc. are the reasons for breast cancer. Two categories of breast cancer are Lobular and Ductal. Ultrasound, MRI, Mammogram are the several diagnosis methods. By employing Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), decision tree (DT), Random Forest (RFA), Naïve Bayes ī (NB), gradient boosting (GB), Logistic regression (LR) and Support Vector Machine (SVM) breast cancer can be predicted. In this paper, we have used the dataset from Kaggle and have applied various ML techniques to after applying dimensionality reduction techniques. The model gives best results when the principal features are selected.

Key Words: *Breast cancer; Logistic regression; Gradient boosting; K-Nearest neighbor; Principal features*

****Corresponding Author**: Gururaj HL, Lecturer, Department of Computer Science & Engineering, Vidyavardhaka College of Engineering, Mysuru, India; E-mail: gururaj1711@vvce.ac.in*

***Received Date**: January 19, 2022, **Accepted Date**: February 22, 2022, **Published Date**: February 28, 2022*

***Citation**: Pavan Kumar SP, Samiha CM, Gururaj HL, Ram Kumar K· Prognostic Analysis of Machine Learning Techniques for Breast Cancer. Int J Bioinfor Intell Comput. 2022;1(1):72-83.*

1. Introduction

Anatomically the tissue of the breast contains liable, alveoli (glands), ducts thatextend from the alveoli to the nipple along with these adipose tissue, connective tissue, lymphatic vessels and lymphatic nodes are found. Breast cancer is initiated due to the uncontrolled proliferation of the cells from any of these breast tissue components. Commonly the lobular tissue components are involved as culprit in most of the breast cancer cases. Breast cancer initiates, develops and enhances with time progress under in-situ condition which later turns to metastatic. Breast cancer may diagnose through self-examination of breast, mammographic screening or based on developed symptoms. Initially the breast cancer will be asymptomatic unless the development of the visible lump in the breast. Most commonly these breast lumps are painless. A significant indicator of the breast cancer malignancy is the deformity of breast, bloody discharge through nipple, breast swelling, heaviness and redness. Since there are very few indications at the early stage of breast cancer development it is best to go through routine screening guidelines for the females to detect any abnormal tissues in the breast. Most of the breast cancer is invasive and the malignancy extends to lymphatic nodes and neighbouring tissues. At present twenty-one classified histologically different breast cancers and four different types of molecular variants are present. These subtypes vary in their overall view and treatment methods [1]. These malignant form exhibits slow growth with aggressiveness. In breast cancer the survival rate of the positive patient depends upon the stage at which diagnosis is done. At an early stage of diagnosis, the survival rate is more (99% in 5-year study of survival rate in the USA) against late stage of diagnosis were 26% at 5-year study of survival rate [2].

2. Why Diagnosis of Breast Cancer is Important? (Epidemiology)

Most commonly caused cancer for women all over the world is breast cancer [3,4]. Every year around 1.7 million are diagnosed positive for breast cancer, which means for every eighteen seconds a new case of breast cancer is reported. In the USA alone 14 % among the total worldwide cases are diagnosed positive [5,6]. Among every 8 positive cases one will be from the USA. Between the age group 55 to 64 are the highest number of cases reported positive for breast cancer. In the molecular level the major gene culprit is Breast Cancer 2 and Breast Cancer 1. These two are the most important genes that are susceptible to breast cancer upon mutation. These genes also induce prostate, pancreatic and ovarian cancer. In the USA one among every 500 may activate for mutation in these genes [7,8].

3. Screening and Detection Methods of Breast Cancer

The appearance of the normal breast tissue was explained and categorized by Stafford Warren [9]. Strickler and Gershon reported the variations of normal and breast cancer tissues through mammography studies. Gershon and et.al., also identified the major characteristic variation in malignant and benign tumors.

3.1 Screening advantages and harms

In the USA during the 1980s to 90s mammography screening of women at the age of 40s led to diagnose early stage of breast cancer, which resulted in the increase of survival rate of positive patients [10]. False positive women may have to undergo breast tissue biopsy tests and other imaging scans for confirmatory positive [11]. After retesting on an average only 10% of the patients were confirmed positive. In film mammography study X-ray is used for imaging records of breast whereas in digital mammography lower radiation doses are used [12]. Breast density study is also noted for mammography as the dense tissue of breast lowers the mammography sensitivity [13].

The constitution of several greasy tissues from breast, but those greasy tissue changes in thick tissue. Internal of this tissue is a system of lobes. Each lobe contains a tiny, tube-like format which is called lobules that accommodate milk organs [9]. Small pipes associate the organs, lobules, and lobes, which transmit milk to the bosom from lobes. The bosom is situated in the areola, which is the black region that encompasses the nipple [10]. Lymph vessels and blood vessels likewise run all over the bosom. Cells will be cherished by the blood. Waste products of the body are depleted from the lymph. A tiny, bean shaped unit and lymph nodes are connected to the lymph vessels and those fights against infection. The lymph vessels interface with lymph hubs, the little, bean-formed organs that help battle contamination [10]. Gatherings of lymph hubs are situated in various territories all through the body, for example, in the neck, crotch, and midsection. Territorial lymph hubs of the bosom are those close by the bosom, for example, the lymph hubs under the arm [10].

4. Literature survey

Breast cancer diagnosis work is carried out with the aid of ANN and ensemble voting algorithm by considering Wisconsin Breast Cancer Database from UCI website [14]. The dataset consists of 699 samples with 30 features. By applying recursive and univariate method, 16 leading features are selected. The dataset is divided into test set and train set. Experimentation is done by applying ANN with logistics algorithm where 15 hidden layers are created with the activation function ReLU [14]. After applying voting algorithm on the outcome, 98.50 % accuracy is obtained.

Accurately classified data are essential to predict or detect the breast cancer. This can be achieved by applying KNN, NB algorithms and SVM [15]. Dataset with 699 samples and 30 attributes are fetched from UCI Machine Learning Repository [15] but from this top 10 features are selected. The implementation is done using Weka software. Firstly, experimentation is carried out for KNN with 10 folds. Best accuracy of 96.85% is obtained when k=3. Secondly, the task is performed with NB and 95.99% of accuracy is achieved. Lastly, SVM classifier is used and when the parameters are $\Upsilon=2^{-15}$ and $C=2^{15}$, accuracy of 96.85% is achieved.

By applying NB, SVM and ANN dataset of breast cancer can be classified [16]. The dataset of 699 instances with 11 attributes are selected from Wisconsin Breast Cancer Diagnosis (WBCD) [16]. Inconstantly achieved values are eliminated and the label column is represented in 0 or 1. 90% of the dataset is used for training and 10% of the dataset is used for testing. The observation of the build model is made by employing ANN, SVM and NB. RBF, Linear, Polynomial and Sigmoid are used as a unit of SVM. SVM Linear Kernel results highest accuracy of 96.72%. RBF NN and Feed Forward NN are applied as a unit of ANN. Radius Basis Function yields accuracy of 95.88%. MultinomialNB, GaussianNB and BernoulliNB are employed as a segment of Naïve Bayes. Accuracy of 95.86% is achieved through GaussianNB.

Diagnosis of breast cancer work is carried out by employing Holo entropy enabled decision tree classifier with Wisconsin dataset [17]. The hole entropy is estimated for every feature occur in the dataset and the attribute which carry the maximum rate of holo entropy [17] is elected just as preeminent attribute. The accuracy is calculated for each chunk size by building DT model. The highest of 99.39% accuracy is obtained.

The formula **for the estimation of Holoo entropy** [4] Where,

$$w = 2*1 - \frac{1}{1+\exp(-E(Ai))} \qquad [1]$$

$$E(Ai) = -\sum_{i=1}^{\mu(Ai)} Pi\ log\ Pi \qquad [2]$$

By employing random forest algorithm types of breast cancer is classified using fine needle aspiration biopsy data [18]. The work is carried out with the size of 699 X 10 datasets where 458 samples are normal case and 241 are infected. Achieved Specificity, Accuracy and Sensitivity are 70%, 72% and 75% respectively for the RF model.

Formula for evaluation metrics,

$$Precision\ = \frac{TrPos}{TrPos + FalPos} \qquad [3]$$

$$Recall\ = \frac{TrPos}{TrNeg + FalNeg} \qquad [4]$$

$$F1\ Score\ = 2 * \frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad [5]$$

$$Accuracy\ = \frac{TrNeg + TrPos}{TrNeg + TrPos + FalPos + FalNeg} \qquad [6]$$

By applying random forest, linear regression and decision tree, classification and prediction of breast cancer work is carried out with the help of Wisconsin breast cancer dataset [19]. Linear regression is applied to the several attributes such as size and shape uniformity, clump thickness etc. Accuracy differs for each attribute. Highest accuracy of 84.15% is achieved. The result obtained from the random forest algorithm is 88.14% [19].

Breast cancer survivability is predicted by applying two machine techniques such as rule induction and random forest [20]. Information on total cancer sufferer in the Gaza strip since 2011 is considered as a dataset from the Palestinian Ministry of Health [20]. The size of the dataset is 1037 X 29. Random forest and rule induction models are used to predict breast cancer with 30% of test set and 70% of the training set. The accuracy achieved in Rule induction model is 73.63% and it took 3.0s to complete the execution [20]. Accuracy of the random forest model is 74.06% and it took 0.5s to build the model.

5. Analysis

Breast cancer is one of the dominant reasons for death in women, so the Anticipation and classification of breast cancer are essential. Based on various features breast cancer can be predicted. Several machine learning techniques such as ANN, SVM, RFA, DT, NB, LR, Rule

Induction and KNN helps us to spot breast cancer. Comparison of various machine learning techniques is shown in Table 1.

Table 1: *Comparison of Machine Learning Techniques.*

Technique	Key Reference	Observation	Accuracy
ANN, NB, SVM ensemble voting algorithm	N. Khuriwal and N. Mishra, 2018.	Out of 30, 16 dominant features are selected by employing recursive and univariate method.	98.50%
SVM, KNN, NB	B. Akbugday, 2019.	Out of 30, 10 dominant features are selected. SVM and KNN mode gave good accuracy compared to NB.	SVM: 96.85% KNN: 96.85% NB: 95.99%
ANN, NB, SVM	M. I. H. Sorrow, M.T. Islam et al., 2019.	The SVM model gave good accuracy compared to ANN and NB when dominant 11 attributes are selected for an experiment.	ANN: 95.88% NB: 95.86% SVM: 96.72%
DT	S. Syed, S. Ahmed and R. Poonia, 2017.	Holo entropy is estimated for individual features and the features which has highest holo entropy value is selected.	99.39%
RFA	F. K. Ahmad and N. Yusoff, 2013.	RFA model achieved good accuracy for 10 Principal attributes. The feature selection method is not discussed.	72%
RFA, LR, DT	S. Murugan, B. M. Kumar and S. Amudha, 2017.	The accuracy achieved through DT model is not mentioned. RFA gave highest accuracy compared to LR when the experimentation is done by choosing different attributes.	RFA: 88.14% LR: 84.15%
Rule Induction, RFA	M. A. M. Alhaj and A. Y. A. Maghari, 2017.	Experimentation is done on a huge dataset with the size 1037 X 29. RFA model achieved better accuracy compared to RI.	RI: 73.63% RFA: 74.06%

By observing varies implementation method we conclude that the best machine learning technique to predict breast cancer is the decision tree. It gives good accuracy of 99.39% when the attributes are selected after estimate Holo entropy values. ANN, SVM and KNN also give a good accuracy when the dominant features are used for the implementation, but in most of the work feature selection method is not discussed.

6. Methodology of Related Work

Performance of the ML models depends on the leading attributes and it is obtained by applying appropriate feature selection method. Feature selection is self-reliant of any machine learning algorithms [21]. By doing this machine learning algorithms can be trained with very less time and an improved accuracy can also be achieved [21].

In this work, we have collected the dataset from the Kaggle which consists of 569 instances and 30 features [22]. The need of dimensionality reduction arises only when the datasethas a greater number of attributes or features because the elementary problem linked with large dimensionality leads to model overfitting. To drain noisy data and to choose dominant features PCA- Principal Component Analysis is used [23]. Steps involved in this work is as shown in Figure 1.

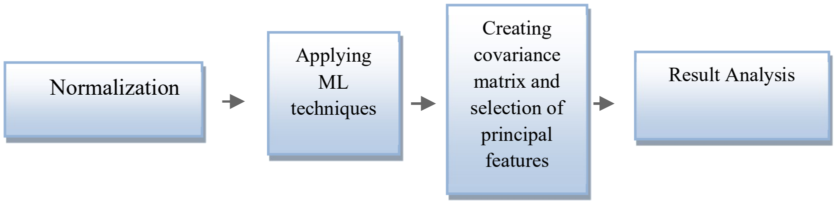
Normalization
Applying ML techniques
Creating covariance matrix and selection of principal features
Result Analysis

Figure 1: *Steps involved in related work.*

6.1 Data normalization

The four goals of data normalization process are, to eliminate duplicate data, to resolve combat data, to configure data and data consolidation [24]. We normalized the dataset by using Z-score functions before applying PCA. This function measures the corresponding Z-score of admission data, relative to standard deviation and sample mean. The key Z-score formula is [25].

$z = (x - \mu) / \sigma$

6.2 Creating covariance matrix and selecting principal features

To the normalized dataset, covariance amid all dimensions will be calculated and it will be stored in the matrix. This matrix shows the relationships among the dimensions. The principal features are selected by noticing the variance ratio of components by employing PCA [26]. Selection of principal feature completely relies on the agreement between loss of information and dimensionality reduction. The dominant 10featurese of a breast cancer dataset based on the covariance are as shown in Figure 2.

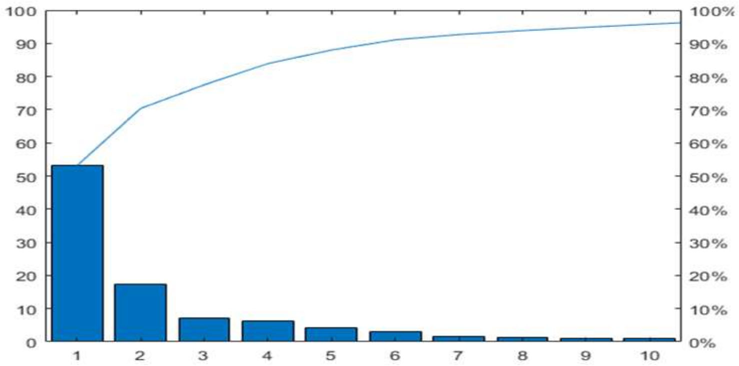

Figure 2: *Dominant 10 features of a breast cancer dataset based on covariance.*

6.3. Applying various machine learning techniques

After performing PCA, the dataset is divided into 80:20 ratio as training and test set. The last column in the dataset represents the status of disease as B or M. B represents Benign and M represents Malignant. Different classification algorithms are applied to the dataset.The result of each machine learning algorithm varies. For the performance evaluation Specificity, Precision, Accuracy, F1 Score, Support and Sensitivity are considered. The experimentation is carried out by considering 30, 20 and 10 features.

6.4. Employed algorithms are

6.4.1. Logistic Regression

One of the popular classification algorithms is Logistic regression. Depend on the set of self-reliant variables discrete values are measured. Another name of logistic regression is logit regression because the possibility of presence of fact is predicted by fitting the dataset to logit function and the result lies within 0 and 1. Logistic Regression (LR) gives highest specificity, accuracy, precision and sensitivity when the top 10 dominant features are selected. The specificity, accuracy, precision and sensitivity obtained from this experiment ranges from 75% to 100%, 93% to 100%, 84% to 100% and 87% to 100% respectively.

6.4.2. Naïve Bayes

The assumption of Naïve Bayes classifier is that the existence of a specific feature or dimension in a class is distinct to the occupancy of any other feature. Gaussian model is built to test the dataset. The model is fitted with the attributes in three sets. Naïve Bayes gives the best specificity, accuracy, precision and sensitivity when the top 10 dominant features are selected. The specificity, accuracy, precision and sensitivity achieved from this experiment ranges from 75% to 100%, 84% to 100%, 84% to 100% and 87% to 100% respectively.

6.4.3. Artificial Neural Network (ANN)

Interconnected unit of nodes from artificial neural network. It has input layer, hidden layers, and output layer. Bias term is added to the weighted sum of all the input layers and weighted sum is called activation. Finally, it is sent through activation function to provide output. We have added 10 hidden layers and the activation function used here is ReLU.

Compared to Logistic regression and Naïve Bayes, Artificial Neural Network (ANN) yields good 97 %accuracy when all the 30 features are used. Selection of principal feature plays vital role in the performance evaluation. Performance of Artificial Neural Network (ANN) with respect to different feature set is as shown in Figure 3.

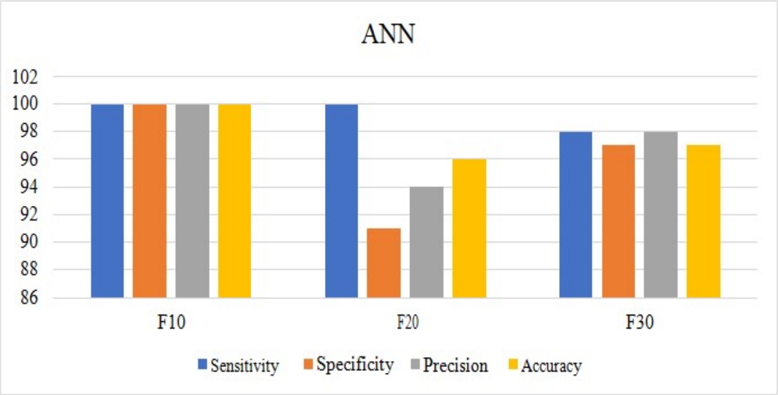

Figure 3: *Performance of Artificial Neural Network (ANN) with respect to different feature set.*

6.4.4. Support Vector Machine

The supervised machine learning algorithm which is employed for both regression and classification is a Support Vector Machine (SVM). The SVM model is built and trained with 80% of the dataset and finally tested with 20% of the dataset. The kernel function is used here is linear. Compared to NB and LT, SVM gave good results. Figure 4 shows the Performance of Support Vector Machine (SVM) with respect to different feature set.

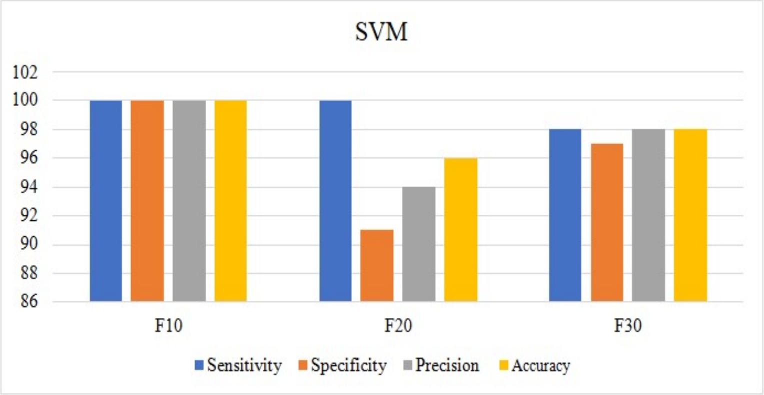

Figure 4: *Performance of Support Vector Machine (SVM) with respect to different feature set.*

6.4.5. Gradient Boosting

XGB-Extreme Gradient Boosting [16] frames object of DT to figure out gradients. In this implementation, from XGBoost library XGBClassifier is imported. The model gave 100% accuracy when dominant 10 features are selected, but an average of 92.5% accuracy is achieved when top 20 and all the features are selected. Accuracy obtained through XGB is great compared to NB but the satisfactory rate of specificity is achieved compared to all other method.

6.4.6. Random Forest Algorithm

Here the Random Forest classifier is used and the predictive model is trained with 80% of the dataset and 20% of the dataset is used for testing. The result achieved through RFA is great when top 10 features are selected. Overall performance of RFA is good. Even when all the features considered, the model gave 95% accuracy. Figure 5 represents the performance of RFA with respect to different feature set.

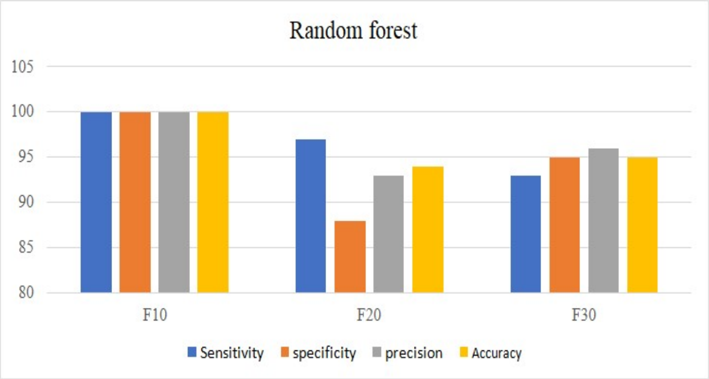

Figure 5) *Performance of Random Forest Algorithm with respect to different feature set*

6.4.7. Decision Tree (DT)

An observation made during the execution of DT is, Precision value 93% remains the same when the top 20 and above features are selected. But there is a huge difference in the accuracy. 88% of accuracy is obtained when 20 features are selected whereas for all the features 93% of accuracy is achieved. Selection of dominant feature plays an important role because best result is achieved when the number of features are 10 Figure 6.

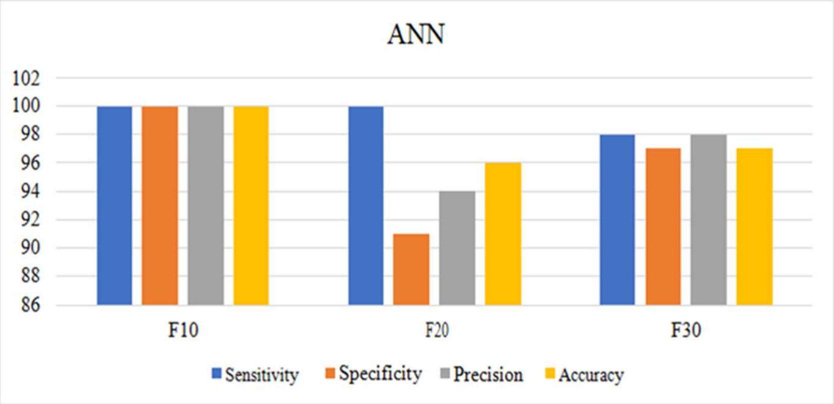

Figure 6: *Represents the accuracy of various machine learning algorithms*

6.4.8. K-Nearest Neighbor (KNN)

Compared to all other method, result achieved through KNN is average. The highest accuracy is achieved when the number of dimensions is 10 but least accuracy is achieved when all the features are selected.

6.5. Result Analysis

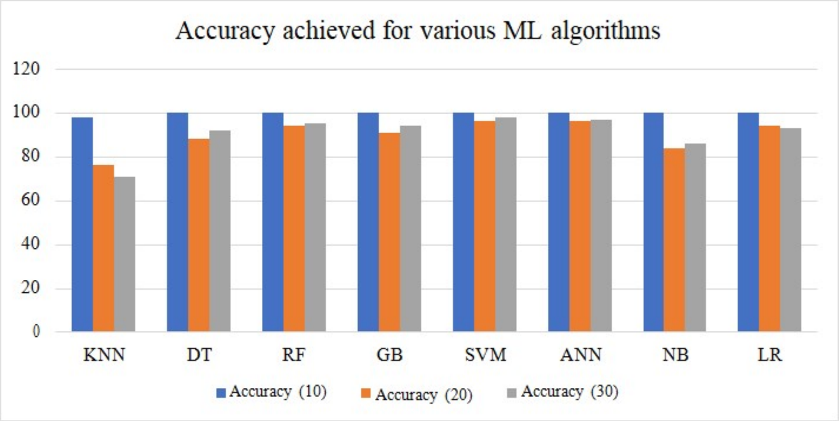

Figure 7: *Accuracy of various machine learning algorithms*

7. Conclusion

Feature selection plays an important role in research work. By applying the appropriate technique, dominant features should be selected. In this work, after data normalization PCA is employed. PCA results features in ascending order based on the covariance value. Various machine learning models such as LR, SVM, XGB, DT, KNN, ANN, RFA and NB are created and results are monitored. The experimentation is done by selecting dominant ten, twenty and all the features. It is observed that the ANN, SVM and RFA gives an accuracy of 97%, 95% and 95% respectively, when all the features are considered, but predictive model gives better accuracy of 100% when the principal ten features are considered.

References

1. Breast Cancer Facts & Figures. American Cancer Society. 2017.

2. PDQ Screening, Prevention Editorial Board. Breast cancer screening (PDQ)-Health Professional Version. J Natl Cancer Inst. 2017.

3. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:359-86.

4. Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN2012, cancer incidence and mortality worldwide: IARC. 2014;1.

5. International Agency for Researchon Cancer. 2013. Available from: http://globocan.iarc.fr

6. DeSantis CE, Fedewa SA, Goding Sauer A, et al. Breast cancer statistics, 2015: convergence of incidence rates between black and white women. Cancer J Clin. 2016;66:31-42.

7. Howlader NNA, Krapcho M, Miller D, et al. Cancer Statistics Review, 1975-2013-SEER Statistics. Based on November 2015 SEER data submission. 2016.

8. https://seer.cancer.gov/csr/1975_2013/

9. Moyer VA. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2014;160:271-81.

10. Rodriguez JL, Thomas CC, Massetti GM, et al. CDC grand rounds: family history and genomics as tools for cancer prevention and control. Morb Mortal Wkly Rep. 2016;65:1291-4.

11. Bassett LW, Gold R. The evolution of mammography. Am J Roentgenol. 1988;150:493-8.

12. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast- cancer incidence. N Engl J Med. 2012;367:1998-2005.

13. Siu AL. Force USPST. Screening for breast cancer: U.S. preventive services task force recommendation statement. Ann Intern Med. 2016;164:279-96.

14. Khuriwal N, Mishra N. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi. 2018.

15. Akbugday B. Classification of Breast Cancer Data Using Machine Learning Algorithms. 2019 Medical Technologies Congress (TIPTEKNO), Turkey. 2019.

16. Showrov MIH, Islam MT, Hossain MD, et al. Performance comparison of three classifiers for the classification of breast cancer dataset. 4th International Conference, Bangladesh. 2019.

17. Sayed S, Ahmed S, Poonia R. Holo entropy enabled decision tree classifier for breast cancer diagnosis using wisconsin (prognostic) data set. 7th International Conference, Nagpur. 2017.

18. Ahmad FK, Yusoff N. Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. 13th International Conference, Bangi. 2013.

19. Murugan Kumar SB, Amudha S. Classification and prediction of breast cancer using linear regression, decision tree and random forest. 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore. 2017.

20. Alhaj MAM, Maghari AYA. Cancer survivability prediction using random forest and rule induction algorithms. 8th International Conference, Jordan. 2017.

21. http://texas-air.org/wp-content/uploads/2021/03/S12-Predictive-Analytics-Process-Using-Machine-Learning-for-Students-Retention.pdf

22. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

23. https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db

24. https://www.statisticshowto.com/probability-and-statistics/z-score/#Whatisazscore

25. https://towardsdatascience.com/detect-parkinsons-with-10-lines-of-code-intro-to-xgboost-51a4bf76b2e6